A7

6/5/1 (Item 1 from file: 347)
DIALOG(R)File 347:JAPIO

04384503 **Image available**
DOCUMENT RETRIEVING DEVICE

PUB. NO.: 06-028403 **JP 6028403** A]
PUBLISHED: February 04, 1994 (19940204)
INVENTOR(s): SASAKI MIKIRO
ISHIKAWA HIROMICHI
APPLICANT(s): MITSUBISHI ELECTRIC CORP [000601] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 04-182438 [JP 92182438]
FILED: July 09, 1992 (19920709)
INTL CLASS: [5] G06F-015/40
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications)
JAPIO KEYWORD:R139 (INFORMATION PROCESSING -- Word Processors)
JOURNAL: Section: P, Section No. 1735, Vol. 18, No. 247, Pg. 87, May 11, 1994 (19940511)

ABSTRACT

PURPOSE: To execute a retrieval by a paragraph unit, and to execute an exact and high speed retrieval in accordance with a purpose by providing a means for discriminating the part corresponding to an item stored in a format storage means and retrieving its index.

CONSTITUTION: As for a document inputted by a document input means 1, in an index preparing means 2, paragraph-constituting information of the document classification concerned is obtained by referring to format data in a format storage means 5, and divided into plural paragraphs. These respective paragraphs are collected at every classification, and prepared as an index. The index generated by the index generating means 2 is stored by an index storage means 3. A fixed length paragraph index storage means 19 can manage index data by data of one table format, because data length of the paragraph is determined. In such a way, in the fixed length paragraph index storage means 19, a relational data base can be realized, and a high speed retrieval can be executed.

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平6－28403

(43)公開日　平成６年(1994)２月４日

(51)Int.Cl.[5]	識別記号	庁内整理番号	Ｆ Ｉ	技術表示箇所
Ｇ０６Ｆ　15/40	５００　Ｔ	7218－5L		

審査請求　未請求　請求項の数４（全 21 頁）

(21)出願番号　特願平4－182438

(22)出願日　平成４年(1992)７月９日

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号

(72)発明者　佐々木　幹郎
鎌倉市大船五丁目１番１号　三菱電機株式会社情報電子研究所内

(72)発明者　石川　浩通
鎌倉市大船五丁目１番１号　三菱電機株式会社情報電子研究所内

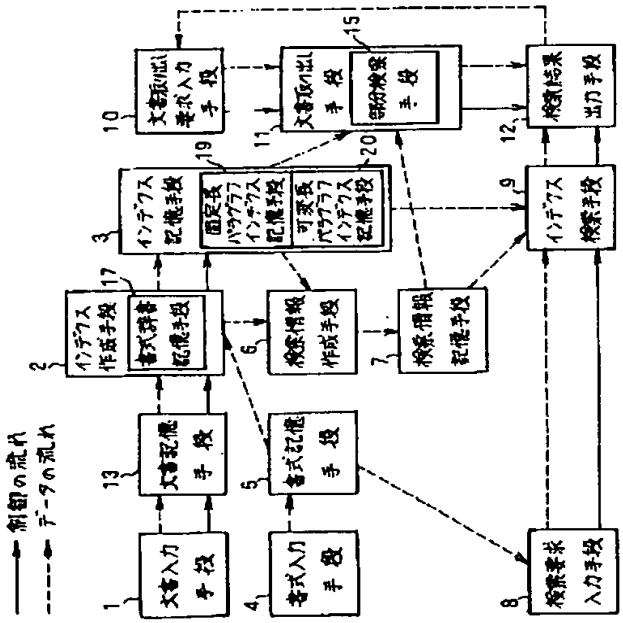
(74)代理人　弁理士　高田　守

(54)【発明の名称】　文書検索装置

(57)【要約】

【目的】　文書の必要な部分のみを高速に検索できる文書検索装置を得る。

【構成】　２は入力された文書をパラグラフに分割してインデクスを作成するインデクス作成手段、３はインデクスを記憶するインデクス記憶手段、５は書式を記憶する書式記憶手段、６はパラグラフの検索の目安となる位置情報を作成する検索情報作成手段、７は検索情報作成手段６により得られた情報を記憶する検索情報記憶手段、１５は文書取り出しを文書単位でなく、パラグラフ単位で行える部分検索手段、１２は検索した文書において、文書単位だけでなく、パラグラフ単位で内容文章をも出力できる検索結果出力手段である。



1

【特許請求の範囲】

【請求項1】　所定の書式を有する文書を検索する文書検索装置において、以下の要素を有する文書検索装置

（a）文書を記憶する文書記憶手段、

（b）文書の書式の構成を示す項目を記憶する書式記憶手段、

（c）文書記憶手段により記憶される文書から、書式記憶手段に記憶されている項目に対応する部分を識別して、その部分を項目に対応するインデクスとして記憶するインデクス記憶手段、

（d）インデクス記憶手段により記憶されるインデクスを検索するインデクス検索手段。

【請求項2】　上記インデクス記憶手段は、項目に対応して、形式の異なるインデクスを有することを特徴とする請求項1記載の文書検索装置。

【請求項3】　所定の書式を有する文書を検索する文書検索装置において、以下の要素を有する文書検索装置

（a）文書を記憶する文書記憶手段、

（b）文書の書式の構成を示す項目を記憶する書式記憶手段、

（c）書式記憶手段により記憶される項目に対応する項目を書式辞書として記憶する書式辞書記憶手段、

（d）文書記憶手段により記憶されている文書と、書式記憶手段に記憶されている項目と、書式辞書記憶手段により記憶される書式辞書に基づき、インデクスを作成して記憶するインデクス記憶手段、

（e）インデクス記憶手段により記憶されるインデクスを検索するインデクス検索手段。

【請求項4】　所定の書式を有する文書を検索する文書検索装置において、以下の要素を有する文書検索装置

（a）文書を記憶する文書記憶手段、

（b）文書の書式の構成を示す項目を記憶する書式記憶手段、

（c）文書記憶手段により記憶される文書から、書式記憶手段に記憶されている項目に対応する文書の部分を識別して、その部分を項目に対応するインデクスとして記憶するインデクス記憶手段、

（d）インデクス記憶手段により記憶されるインデクスに対応する文書の部分を検索するための検索情報を記憶する検索情報記憶手段、

（e）検索情報記憶手段により記憶される検索情報を用いて、文書の部分を検索する部分検索手段。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】この発明は蓄積された文書の中からキーワードを用いて、特に書式が予め決まっている文書を検索する文書の検索装置に関するものである。

【０００２】

【従来の技術】図２２は第41回情報処理学会全国大会で発表されたフルテキストデータベース検索システム「検

2

蔵君」の文書検索装置を示すブロック構成図である。なお、図中実線は制御の流れを示し、点線はデータの流れを示している。図において、１は文書を入力するワープロなどの文書入力手段であり、入力された文書は磁気ディスク装置などの文書記憶手段１３に格納される。８は検索要求として検索条件となるキーワードなどを入力する検索要求入力手段、１４は入力されたキーワードに基づき文書記憶手段内の文書データを検索し、検索要求に適合する文書を限定する文書検索手段、１０は使用者がどの文書を取り出すのかを指定する文書取り出し要求入力手段、１１は文書取り出し要求に基づいて文書記憶手段１３から対象文書を取り出す文書取り出し手段、１２はどの文書が検索要求に適合したかという検索結果や取り出した文書データなどをディスプレイなどの出力装置に示す検索結果出力手段である。

【０００３】従来例の動作について、図２２のブロック図を用いて説明する。文書そのものはワープロなどの文書入力手段１によって入力され、文書記憶手段１３に記憶される。

【０００４】従来例の検索処理について説明する。文書内容との照合を行うためのキーワードなどの検索要求が検索要求入力手段８により入力される。入力された検索要求は、文書検索手段１４へ送られ、文書検索手段１４により文書記憶手段１３内のすべての文書内容についてキーワードによる検索が行われる。検索の結果、検索要求に適合する文書が限定されるとそれらの文書情報は検索結果出力手段１２へ送られ、検索結果出力手段１２によりディスプレイなどに出力される。

【０００５】次に文書の取り出し処理について説明する。使用者は出力された検索結果を見て取り出したい文書を決め、文書取り出し要求手段１０により文書番号や文書名などの文書取り出し要求を入力する。入力された文書取り出し要求は文書取り出し手段１１に送られ、文書記憶手段１３の中から文書取り出し要求に一致する文書のデータが文書取り出し手段１１により取り出され、検索結果出力手段１２へ送られる。検索結果出力手段１２は送られてきた文書データをディスプレイなどに出力する。

【０００６】他の従来例について説明する。図２３は、例えば、（株）平和情報センターの日本語文書情報系データベースシステム" Future/Happiness" のカタログに記載されているような従来のキーワードを用いた文書検索装置を示すブロック構成図である。なお、図中、実線は制御の流れを示し、点線はデータの流れを示している。

【０００７】図２３を参照しながらこの従来例の構成を説明する。図において、１は文書を入力するワープロ等の文書入力手段であり、入力された文書は磁気ディスク装置等の文書記憶手段に格納される。２３１は格納された文書から自然言語処理機能によりキーワードを自動的

に抽出登録するキーワード登録手段であり、抽出されたキーワードは磁気ディスク装置等のキーワード記憶手段に格納される。8は検索要求となるキーワード等を入力するキーボードディスプレイ等の検索要求入力手段、14は入力されたキーワードに基づきキーワード記憶手段16を用いて検索要求に適合する文書を限定する文書検索手段、11はこの限定された文書の実体を、上記文書記憶手段13より取り出す文書取り出し手段であり、このようにして得られた検索結果はディスプレイ等の検索結果出力手段12により出力される。

【0008】次に動作について説明する。先ず文書そのものは文書入力手段1により入力され、文書記憶手段13に格納される。また同時に、キーワード登録手段231では入力された文書を文法に基づき、単語毎に分割する分かち書き処理を行った後、分割された単語を評価して、助詞等の不要な用語を除去することにより、検索時に必要となるキーワードを自動的に抽出し、キーワード記憶手段232に格納する。

【0009】次に、検索時について説明する。検索者は文書記憶手段13に記憶されている文書の中から自分の欲しい文書を検索する場合、検索要求入力手段8に検索条件としてキーワードを入力する。文書検索手段14は、キーワード記憶手段232に格納されているキーワードと、検索要求入力手段8で入力されたキーワードとのマッチングを行い、検索者の欲する文書を検索する。文書取り出し手段11では、文書検索手段14で検索された文書の実体を文書記憶手段13から取り出し、検索結果出力手段12であるディスプレイ等の表示装置に出力する。

【0010】この従来例の文書検索装置は以上のようにして、蓄積された大量の文書からキーワードをもとにして検索を行うものである。

【0011】

【発明が解決しようとする課題】従来の文書検索装置は以上のように構成されているので、キーワードなどの検索要求により、全文書の全範囲を検索する必要があり、検索に時間がかかるという問題点があった。また、入力されたキーワードをもとに全文書にわたって検索されるために、検索者が本来期待しない文書部分でのマッチングが行われ、不要な文書まで出力されてしまうという問題点があった。さらに、検索結果として文書を取り出す場合、文書単位でしか取り出せず、検索者は、その文書の中から自分の見たい必要なところを探さなければならなかった。

【0012】また、従来の文書検索装置は、以上のように構成されていたので、文書データに対して付加したキーワードを基に検索を行うことはできても、このキーワードが数および種別を限定されているために、キーワードで文書の内容を詳細に表現することは困難であり、文書の内容を示す詳細な事項を指定して文書の検索を行うことはできないという問題点があった。また、キーワードを基に検索を行う場合、上記に示したようにキーワードでは多様な意味を持つ文書の内容をすべて表現することはできないため、検索者が欲したものの他に、不要なものまで検索されてしまうという問題点があった。

【0013】この発明はこのような問題点を解消するためになされたものであり、高速に文書の検索ができ、文書中の必要な部分だけを取り出せる文書検索装置を得ることを目的とする。

【0014】

【課題を解決するための手段】この発明の請求項1に係る文書検索装置は、検索すべき対象となる文書を記憶する文書記憶手段と、検索する文書の書式がどのような構成であるかを示す複数の項目を記憶する書式記憶手段と、文書記憶手段に記憶されている文書の内容を見て、書式記憶手段に記憶されている複数の項目に対応して文書の部分が、文書の中にあるかどうかを識別し、項目に対応する部分があったなら、その部分を項目に対応するインデクスとして記憶するインデクス記憶手段と、インデクス記憶手段により記憶されるインデクスの中に存在する検索すべき対象となる部分を検索するインデクス検索手段を備えたものである。

【0015】この発明の請求項2に係る文書検索装置は、請求項1のインデクス記憶手段において、項目の種別に対応して、形式の異なる複数種類のインデクスを有することができるものである。

【0016】この発明の請求項3に係る文書検索装置は、検索すべき対象となる文書を記憶する文書記憶手段と、検索する文書の書式がどのような構成であるかを示す複数の項目を記憶する書式記憶手段と、書式記憶手段により記憶される項目を特定の群として定める書式辞書を設けて、それを記憶する書式辞書記憶手段と、文書記憶手段に記憶されている文書の内容と、書式記憶手段により記憶されている書式辞書とに基づき、書式記憶手段の項目を書式辞書の項目と比べ、書式に項目に対応する部分が文書にあるかどうかを識別し、それにより、文書の部分を項目に対応するインデクスを作成して、記憶するインデクス記憶手段と、インデクス記憶手段により記憶されるインデクスの中に存在する検索すべき対象となる部分を検索するインデクス検索手段を備えたものである。

【0017】この発明の請求項4に係る文書検索装置は、検索すべき対象となる文書を記憶する文書記憶手段と、検索する文書の書式がどのような構成であるかを示す複数の項目を記憶する書式記憶手段と、文書記憶手段に記憶されている文書の内容を見て、書式記憶手段に記憶されている複数の項目に対応して文書の部分が、文書の中にあるかどうかを識別し、項目に対応する部分があったなら、その部分を項目に対応するインデクスとして記憶するインデクス記憶手段と、インデクス記憶手段に

5

より記憶される文書のある部分を検索するために、インデクスにおける文書の部分が、文書記憶手段に記憶されている文書のどの部分に存在するかという位置情報を記憶している検索情報記憶手段と、検索情報記憶手段により記憶される位置情報を用いて、文書記憶手段により記憶される文書の部分を検索する部分検索手段を備えたものである。

【００１８】

【作用】この発明の請求項１の文書検索装置においては、文書が一定の書式に従って書かれている点に着目し、インデクス記憶手段はあらかじめ入力しておいた文書書式を用いて、文書データを論理的にまとまりのあるパラグラフに分割し、インデクスとして格納し、このインデクスを検索することにより、文書全体を検索の対象とせず、パラグラフ単位で検索ができる。また、検索時に、このようなインデクスを用いることにより、パラグラフを検索することになり、文書の内容に関する詳細な事項を指定して検索を行うことができる。

【００１９】この発明の請求項２の文書検索装置においては、インデクス記憶手段は、インデクスの作成時、例えば、特許明細書における「書類名」、「発明の名称」などの構成を示す項目において、その項目に記される内容文章の長さが定まるパラグラフについては、固定長であるパラグラフとして、その他長さが定まらないものについては、可変長であるパラグラフとして、格納することができるようにインデクスの種類を分けるので、検索要求に対して、どちらかの適切なインデクスを選択してパラグラフの検索を行うことができる。

【００２０】この発明の請求項３の文書検索装置においては、辞書を設けた、書式内に含まれない任意の項目を辞書に登録しておくことにより、書式内の項目と同様に扱い、それに従い、インデクスを検索できる。

【００２１】この発明の請求項４の文書検索装置においては、検索情報記憶手段の記憶する位置情報に従い、部分検索手段により、検索要求に適合した文書に関して適合したパラグラフのみを検索結果として文書記憶手段から取り出すことができるので、必要な情報のみ得ることができる。

【００２２】

【実施例】

実施例１．図１は、実施例１におけるの文書検索装置の全体構成を示すブロック図である。従来例の図２２と同一または相当部分には同一符号を用い、その説明は省略する。図１において、２は入力された文書を書式に基づいて、論理的な意味を持つ文章のかたまりであるパラグラフに分割して、パラグラフの種別毎にまとめたインデクスを作成するインデクス作成手段である。３はインデクス作成手段２により、パラグラフの種別ごとにまとめたインデクスを記憶する磁気ディスクなどのインデクス記憶手段である。４は、１つの文書がどのようなパラグラフで構成されているのかを示す書式を、文書の種別ごとに入力する書式入力手段、５は、その書式を記憶する磁気ディスク装置などの書式記憶手段である。６はインデクス作成手段２によって分割されたパラグラフが、文書記憶手段１３に記憶されている元の文書においてどの部分だったのか、そして、インデクス記憶手段３内のどの場所に記憶されているのかというパラグラフの検索の目安となる位置情報を作成する検索情報作成手段、７は検索情報作成手段６により得られた情報を記憶する磁気ディスク装置などの検索情報記憶手段である。８は前記図２２で示した検索要求入力手段８を改良したもので、検索要求を１つの文書全体に対してだけでなく、１つの文書内の各パラグラフに対しても出すことができるようになっている。９は検索要求手段８で入力された検索要求に対し、インデクス記憶手段３内のインデクスのパラグラフを検索し、１パラグラフの検索ごとに検索範囲の絞り込みを行うインデクス検索手段である。１０は前記図２２で示した文書取り出し要求入力手段１０を改良したもので、文書取り出し要求を文書単位でなく、文書内のパラグラフ単位で指定できるようにしたものである。１１は前記図２２で示した文書取り出し手段１１を改良したもので、文書取り出し要求入力手段１０の指定により、文書取り出しを文書単位でなく、パラグラフ単位で行えるようにした部分検索手段１５を含んでいる。１２は前記図２２で示した検索結果出力手段１２を改良したもので、文書取り出し手段１１により取り出してきたデータを、文書単位だけでなく、パラグラフ単位でも出力できるようにしている。なお、新たに備えられたインデクス作成手段２や検索情報記憶手段７などは、計算機システムを構成するプロセッサとその上で動作するソフトウェアによって実現されている。

【００２３】次に動作について説明する。前述したように、本実施例は文書が一定の書式に従って書かれている点に着目したものであり、例えば、規格書・仕様書などの技術文書は図２に示すようにそれぞれ定まった書式を有している。文書の書式は予め書式入力手段４から入力され、書式記憶手段５に格納される。このとき、文書の書式は、書式の同じ文書を扱う場合には１種類でよいが、種別の異なる文書を扱う場合には文書の種別によって書式が異なるために、その種別の数だけ入力される。これらの文書の種別や構成を示す書式データは、文書をパラグラフごとに分割する際や検索要求入力時に、その種別が、決められた指定を用いて行われる。

【００２４】一般に文書は、パラグラフ（ある意味に従い、かたまりとみなされる論理的な単位）の集まりからなる。このパラグラフには図３の例でいうと、「文書名」や「適用範囲」等のように、その内容文章の長さが何文字以内と定まっているもの（固定長パラグラフ）と、「一般要求事項」のようにその長さが何文字以内というように定まらないもの（可変長パラグラフ）があ

る。従って、書式として入力される書式データは、図４に示すように、文書がどのようなパラグラフから構成されているかを表す文書構造４１、パラグラフの内容を表す書式項目名４２、書式項目名に対して付けられた書式項目番号４３、パラグラフが固定長か可変長かを表す固定長／可変長フラグ４４、書式項目名に対して付けられるパラグラフＩＤ４５などにより構成されることになり、パラグラフ構成情報として対応するフラグで表現される。

【００２５】文書そのものは従来例通り文書入力手段１によって入力される。入力された文書は、インデクス作成手段２において、書式記憶手段５内の書式データの参照により該当する文書種別のパラグラフ構成情報が得られ、複数のパラグラフへと分割される。更に、分割された各パラグラフは、図５のようにパラグラフの種別ごとにまとめられ、インデクスとして作成される。そして、インデクス作成手段２で作成されたインデクスは、インデクス記憶手段３により格納される。

【００２６】次に、本実施例でのインデクスの詳細について説明する。図６は固定長パラグラフのインデクスを記憶する固定長パラグラフインデクス記憶手段の構造を示したものである。この固定長パラグラフインデクス記憶手段については、パラグラフのデータ長が定まるため、インデクスデータを１つの表形式のデータで管理することができる。このため固定長パラグラフインデクス記憶手段では、リレーショナルデータベースの実現が図れ、検索時に高速な検索を行うことができる。

【００２７】図７は可変長パラグラフのインデクスを記憶する可変長パラグラフインデクス記憶手段の構造を示したものである。この可変長パラグラフインデクス記憶手段については、パラグラフのデータ長が定まらないため、各パラグラフのインデクスごとにデータを一つの表形式で管理する。

【００２８】インデクス作成手段２により作成されるインデクスにおいてのパラグラフ分割の前の元文書に関する情報（文書名、作成者など）と、インデクス記憶手段３により記憶されるパラグラフ分割後に関する情報（文書中の位置情報、格納場所など）はリンクされ、図８に示すように、位置情報として検索情報作成手段６により作成され、検索情報記憶手段７に記憶される。

【００２９】次に本実施例の検索処理について説明する。従来の検索要求入力は文書記憶手段１３に登録されている全文書に対し、それぞれの文書の内容すべてを検索対象として、キーワードを入力することで検索が行われていた。本実施例では、図９に示すように文書の種類を選択し、書式記憶手段５内の書式データのパラグラフ構成情報を参照することにより、各パラグラフを検索対象としてキーワードを入力する。入力されたキーワードは、文書の種別の情報や検索対象のパラグラフの種別の情報と一緒にインデクス検索手段９へと送られる。インデクス検索手段９では、検索要求手段８から送られてきた検索要求に基づいて、インデクス記憶手段３に記憶されたパラグラフをキーワードに基づき検索する。その結果、検索要求に適合する文書が限定される。

【００３０】なお、定型な文書書式をもつ文書においては、文書構造、書式項目名、書式項目番号が図４に示したように書式データのパラグラフ構成情報として定まっているが、実際には図１０に示すように、書き手が自分の使いやすいように項目名を変え、書式を変えて文書を作成するケースがある。この様な場合には、同じ書式であっても、違う書式データを作成することになってしまう。本実施例の文書検索装置は、書き手によって文書書式が違う文書に対しても、インデクス作成手段２において、書式辞書及び書式辞書記憶手段を設けたので、その違いに対応することができる。書式辞書とは、定型な文書書式に対して、書き手が自分の作り易いように文書書式を変更して作成した文書からでも、インデクスの作成を可能にするために設けられたものである。具体的には図１１に示すように、予め定まっている書式項目名に対して、書き手が変更して書き得る可能性のある項目名を辞書としたものであり、図４のパラグラフ構成情報のパラグラフＩＤ４５を書式項目名に対応して蓄積している。

【００３１】図１２はインデックス作成手段２の構成を示すブロック図である。１６は書式記憶手段５に記憶されている文書書式の各項目に対応した書式項目名をあらかじめ辞書として入力する、ワープロ等の書式辞書入力手段、１７は入力された書式辞書を記憶する磁気ディスク装置等の書式辞書記憶手段、１８は文書記憶手段１３に蓄積されている文書データの中からその中に書かれている書式項目名を抽出し、抽出された書式項目名と書式辞書記憶手段１７に記憶されている書式辞書の項目名とを比較することによって、分割されたパラグラフと文書書式との対応付けを行うインデクス作成手段である。なお、インデクス作成手段２は、計算機システムを構成するプロセッサとその上で動作するソフトウエアによって実現されている。

【００３２】インデクス作成の動作について図１２を用いて説明する。インデクス作成手段２において、文書入力手段１から文書が入力され文書記憶手段１３に格納される度に、書式記憶手段５に記憶されている文書の書式データを用いて、文書記憶手段１３の文書データをパラグラフに分割し、インデクスを作成する。

【００３３】上記のインデクス作成手段２は、図１３に示すような一連の処理を行う。以下に、それぞれの処理について図１４の文書例をもとに説明を行う。

【００３４】（ａ）書式項目名抽出処理１３１：文書記憶手段１３に蓄積されている文書データの中から、新たに登録された文書データを先頭行から読みだし、書式項目番号（図４の４３参照）を検出した場合は、その次に

書かれている書式項目名を抽出する。図１４の文書では、先ず先頭行として「１．文書名称」が読み込まれ書式項目番号「１．」が検出され、続く「文書名称」が書式項目名として抽出される。

【００３５】（ｂ）パラグラフＩＤ獲得処理１３２；（ａ）で抽出された書式項目名をキーとして書式辞書の検索を行い、マッチングしたデータのパラグラフＩＤを得る。図１１に示した書式辞書を例にとると、図１４の文書では、書式項目名「文書名称」はパラグラフＩＤが「１」に対応することが分かる。

【００３６】（ｃ）パラグラフ抽出処理１３３；（ａ）で書式項目名が検出された次の行から、また書式項目名が検出されるまでデータを読み込む。この際読み込まれたデータがパラグラフである。図１４の文書では、次の書式項目番号「適用する範囲」が検出されるまでのデータ、すなわち、２行目の「△△△仕様書」がパラグラフデータとして抽出される。

【００３７】（ｄ）インデックス作成・登録処理１３４：（ｂ）で得られたパラグラフＩＤおよび、（ｃ）で抽出されたパラグラフデータを図１５に示すように対応付けてインデクスデータとしてインデクス記憶手段に出力する。なお、本実施例では、検索の際に高速化を図るため、先に述べたように固定長パラグラフと可変長パラグラフの場合でインデクスの記憶手段を分けている。インデクスの登録の際には、書式記憶手段５に記憶される書式データのパラグラフ構成情報（文書構造、書式項目名、書式項目番号、パラグラフＩＤ、固定長／可変長フラグ）を参照し、（ｂ）で得られたパラグラフＩＤをキーとして、該当する書式項目の固定長／可変長フラグをチェックし、（ｃ）で得られたパラグラフを固定長パラグラフインデクス記憶手段に登録するか、可変長パラグラフインデクス記憶手段に登録するかを判定している。

【００３８】（ｅ）継続行の判定１３５：（ａ）～（ｄ）の処理の後に、文書内に継続行がある場合は、継続行がなくなるまで繰り返して処理する。

【００３９】次に検索時について説明する。例えば、図１６（ａ）、（ｂ）に示すようなインデクスデータが、上述したインデクス作成手段２によって作成され、それぞれ固定長パラグラフインデクス、可変長パラグラフインデクスに格納される。また図１６（ｃ）に示すような書式データが書式記憶手段５に格納される。このような場合に、検索者が「書式項目『適用範囲』の中に『文書検索装置』というワードがあるもの」というように文書の内容を指定して検索を行いたいときは、検索要求入力手段８において、図１７（ａ）に示すように書式記憶手段５に記憶されている書式を画面上に表示し、検索者はシステムが表示したこの書式の書式項目毎に、文書の内容を表すワードをキーワードとして入力することにより検索が行える。上記の例では、書式項目『適用範囲』に対して、キーワード『文書検索装置』を入力することになる。このようにして入力された、文書の内容を指定した検索条件は、図１７（ｂ）に示すように、書式データのパラグラフＩＤとキーワードを組として、文書名が検索結果として（ｃ）のように検索結果手段１２に出力される。なお、この際、同時に、書式データのうち指定された書式項目の固定長／可変長フラグも同時に出力される。

【００４０】ここで、文書取り出し手段１１は、インデクス記憶手段３に記憶されたインデクスデータを用いて、検索要求入力手段８より入力された検索条件をもとに検索を行う。図１７の例では、先ず、検索条件の中でキーワードが設定された書式項目の固定長／可変長フラグを判定する。図１７の例では検索条件の中で、固定長パラグラフを表すフラグが示されているので、固定長パラグラフインデクスを用いて検索を行う。具体的には、固定長パラグラフインデクスの中で、パラグラフＩＤが「２」の列に存在するパラグラフの中に、キーワードとして入力したワードが存在するかどうかを調べ、存在した場合には検索条件に適合したものとみなす。この場合、図１７（ｃ）に示すように文書Ｂが検索結果として得られるが、さらに、文書取り出し手段１１により、文書記憶手段１３から文書の内容が取り出され、検索結果出力手段８であるディスプレイ等の表示装置に出力される。

【００４１】また、上記では固定長パラグラフに対応した書式項目にキーワードが設定された場合について説明したが、可変長パラグラフに対応した書式項目にキーワードを設定することももちろん可能である。図１８は上記のような検索の具体例を示したものであり、可変長パラグラフに対応した書式項目『一般要求事項』の中に『マルチメディア』というワードがある文書を検索する場合を示している。検索要求の入力は、図１８（ａ）に示すように、書式項目が固定長パラグラフに対応したものと同様に、書式項目『一般要求事項』に対して、キーワード『マルチメディア』を入力するだけでよい。従って、検索者はキーワードを入力しようとしている書式項目が固定長か可変長かについては意識する必要がない。図１８（ｂ）は検索条件を示している。この場合、固定長／可変長フラグが可変長に設定される。

【００４２】ここで、文書取り出し手段１１は、インデクス記憶手段に記憶されたインデクスデータを用いて、検索要求入力手段８より入力された検索条件をもとに検索を行う。この場合、可変長パラグラフを表すフラグが示されているので、可変長パラグラフインデクスを用いて検索を行う。具体的には、可変長パラグラフインデクスデータの中でパラグラフＩＤが「４」のパラグラフデータの中にキーワードとして入力したワードが存在するかどうかを調べ、存在した場合には検索条件に適合したものとみなす。この場合、図１８（ｃ）に示すように文書Ａ、Ｂが検索結果として得られる。以上の検索により

検索結果として得られた文書は、文書取り出し手段１１により文書記憶手段１３より取り出され、検索結果出力手段１２であるディスプレイ等の表示装置に出力される。

【００４３】本実施例での部分検索手段１５について説明する。検索者は出力された検索結果を見て、自分の取り出したい文書を決め、文書取り出し要求手段１０により、文書取り出し要求を入力する。このとき、従来方式では文書単位でしか内容の取り出しが指定できなかったのに対し、本実施例では図１９に示すように、文書のパラグラフまで指定して取り出し要求が入力できる。入力された文書取り出し要求は文書取り出し手段１１に送られ、要求に適合した文書が取り出される。このとき、文書取り出し手段１１は図８に示す検索情報記憶手段７内の位置情報を参照して、部分検索手段１５により、文書記憶手段１３内の文書の内容をパラグラフ単位で取り出すことができる。部分検索手段１５により得られたパラグラフは、検索結果出力手段１２により出力される。

【００４４】実施例２．上記実施例では、一つの書式項目に対してキーワードを入力する場合を説明したが、複数の書式項目に対してキーワードを入力し、検索することももちろん可能である。図２０は上記のような検索の具体例を示したものであり、書式項目『適応範囲』の中に『装置』というワードがあり、しかも、書式項目『一般要求事項』の中に『マルチメディア』というワードがある文書を検索する場合を示している。この場合は、固定長パラグラフインデクス記憶手段１９と可変長パラグラフインデクス記憶手段２０を用いることにより検索が行なわれる。

【００４５】実施例３．実施例１の検索要求入力手段８において、複数のパラグラフを検索対象としてキーワードの入力が行われた場合、インデクス検索手段９は、検索情報記憶手段７を参照して、１パラグラフ検索ごとにインデクス内における検索範囲の絞り込みを行いながら検索を行うことが可能である。図２１のようにインデクスＡ、Ｂ、Ｃに対して結合条件ＡＮＤでそれぞれのパラグラフに対してあるキーワードが入力された場合、インデクス検索手段９はまずインデクスＡの検索を行う。インデクスＡに対する検索の結果、文書１・文書３・文書４が条件に適合したとする。次のインデクスＢの検索を行う際に、インデクス検索手段９は検索情報記憶手段７を参照して、インデクスＢ内の文書１・文書３・文書４の部分のみを検索対象として取り出し、検索を行う。以降はこの繰り返しにより検索範囲の絞り込みが行われる。なお、検索要求の結合条件がＯＲの場合は、条件に適合した文書以外の文書のパラグラフを次の検索対象とすることにより、絞り込みが行われる。そして、インデクス検索手段９により限定された文書のパラグラフは、検索結果出力手段１２によりディスプレイなどに出力される。このように、検索要求が複数のパラグラフに対してある場合、パラグラフごとに検索が行われる際に、前のパラグラフの検索結果から検索要求に適合する文書が限定されるので、検索情報記憶手段を参照することで、その次のパラグラフ検索では前回の検索時より範囲を限定した検索を行うことができる。

【００４６】実施例４．実施例１では、インデクス記憶手段において、文書の分割されたパラグラフを有するとしたが、これは、文書記憶手段に対して文書検索を行うよりも、インデクス記憶手段に対して検索を行うほうが、より高速に検索できるためである。しかし、この方法では、記憶容量の問題があると思われる場合には、インデクス記憶手段においてパラグラフを有さず、そのかわり、パラグラフの存在する文書記憶手段内の文書の位置を情報としてインデクス記憶手段で有することで、インデクス記憶手段のパラグラフの位置情報を元に、常に文書記憶手段に対して文書を検索する方法を取ってもよい（ただし、この方法では、時間的問題がある）。また、インデクス記憶手段では、文書記憶手段に記憶されている文書のパラグラフそのものと全く同じものを記憶していてもよいし、あるいは、文書記憶手段に記憶されている文書のパラグラフにおいて、罫線情報や特殊な文字コード情報などの不要な情報を除いた部分のみを記憶しているものであってもよい。

【００４７】実施例５．実施例１において、インデクス記憶手段では、検索の高速化を狙い、固定長パラグラフインデクス記憶手段と可変長パラグラフインデクス記憶手段とを分別して設けたが、特にこの２種類に分別する必要はなく、例えば、英文と和文の区別と言うように、他の種別による記憶手段を、インデクス記憶手段において、設けるものであってもよい。また。記憶手段の分別は、２種類以上であってもよい。

【００４８】実施例６．実施例１において、書式辞書記憶手段は、インデクス作成手段が有するとして説明したが、特にインデクス作成手段が有さずに、独立した機能として存在してもよい。書式辞書入力手段においても、同様である。

【００４９】実施例７．実施例１において、書式記憶手段に記憶される書式データのパラグラフ構成情報は、図４に示すように、数字のフラグを使用しているが、例えば、文字や記号を使用するものでもよい。また、書式記憶手段に記憶される書式データの構成は、図４のような階層構造以外に、表形式などの形式でも構わない。

【００５０】実施例８．実施例１において、部分検索手段は、文書取り出し手段に含まれるとしたが、独立した機能として存在してもよい。また、部分検索手段の実施方法は図１９に示したが、検索対象とするパラグラフの選択の方法に関しては、特に限定はしない。

【００５１】

【発明の効果】以上のように請求項１の発明による文書検索装置では、文書を記憶する文書記憶手段と、文書の

書式の構成を示す項目を記憶する書式記憶手段と、文書記憶手段により記憶される文書から、書式記憶手段に記憶されている項目に対応する部分を識別して、その部分を項目に対応するインデクスとして記憶するインデクス記憶手段と、インデクス記憶手段により記憶されるインデクスを検索するインデクス検索手段とを設けたことにより、文書の検索時にパラグラフ単位で検索ができるので、検索者の目的に応じて、適格で高速な検索ができる。

【００５２】以上のように請求項２の発明による文書検索装置では、請求項１のインデクス記憶手段が、項目に対応して、形式の異なるインデクスを有することにより、検索対象のパラグラフを固定長と可変長に分別して管理できるので、検索者の必要な文書のみを高速に検索することが可能である。

【００５３】以上のように請求項３の発明による文書検索装置では、文書を記憶する文書記憶手段と、文書の書式の構成を示す項目を記憶する書式記憶手段と、書式記憶手段により記憶される項目に対応する項目を書式辞書として記憶する書式辞書記憶手段と、文書記憶手段により記憶されている文書と、書式記憶手段に記憶されている項目と、書式辞書記憶手段により記憶される書式辞書に基づき、インデクスを作成して記憶するインデクス記憶手段と、インデクス記憶手段により記憶されるインデクスを検索するインデクス検索手段とを設けたことにより、書式の類似している文書を同一書式とみなして検索できるので、検索者の目的に応じた文書を適格に検索できる。

【００５４】以上のように請求項４の発明による文書検索装置では、文書を記憶する文書記憶手段と、文書の書式の構成を示す項目を記憶する書式記憶手段と、文書記憶手段により記憶される文書から、書式記憶手段に記憶されている項目に対応する文書の部分を識別して、その部分を項目に対応するインデクスとして記憶するインデクス記憶手段と、インデクス記憶手段により記憶されるインデクスに対応する文書の部分を検索するための検索情報を記憶する検索情報記憶手段と、検索情報記憶手段により記憶される検索情報を用いて、文書の部分を検索する部分検索手段とを設けたことにより、検索対象の文書を、文書の内容全体としてだけでなく、そのないようの一部分であるパラグラフも検索できるので、検索者の必要な情報のみを高速に得ることができる。

【図面の簡単な説明】

【図１】本発明の実施例１における文書検索装置の全体構成を示すブロック図である。

【図２】本発明の実施例１で用いられる一定の書式に従って書かれている文書の一例を示す図である。

【図３】本発明の実施例１の書式と文書データにおけるパラグラフの対応を示す図である。

【図４】本発明の実施例１の書式データにおけるパラグラフ構成情報の例を示す図である。

【図５】本発明の実施例１のインデクス作成手段により分割された文書のパラグラフが、インデクス記憶手段により格納される例を示す図である。

【図６】本発明の実施例１の固定長パラグラフインデクス記憶手段の構成を示す図である。

【図７】本発明の実施例１の可変長パラグラフインデクス記憶手段の構成を示す図である。

【図８】本発明の実施例１のインデクス記憶手段と検索情報記憶手段の関係を示す図である。

【図９】本発明の実施例１の検索処理の流れを示す図である。

【図１０】本発明の実施例１の文書書式の一例を示す図である。

【図１１】本発明の実施例１の文書書式と書式辞書の対応を示す図である。

【図１２】本発明の実施例１のインデクス作成手段の構成を示す図である。

【図１３】本発明の実施例１のインデクス作成手段の処理の流れを示す図である。

【図１４】本発明の実施例１の文書例を示す図である。

【図１５】本発明の実施例１の文書データと書式データにおけるパラグラフＩＤの対応を示す図である。

【図１６】本発明の実施例１の固定長パラグラフインデクスデータと可変長パラグラフインデクスデータにおけるパラグラフＩＤの書式データへの対応を示す図である。

【図１７】本発明の実施例１の固定長パラグラフに対応した検索方法の一例を示す図である。

【図１８】本発明の実施例１の可変長パラグラフに対応した検索方法の一例を示す図である。

【図１９】本発明の実施例１の部分検索手段によるパラグラフの検索方法を示す図である。

【図２０】本発明の実施例２の検索方法を示す図である。

【図２１】本発明の実施例３の検索においてインデクスのパラグラフの絞り込み方法を示す図である。

【図２２】従来例の文書検索装置の全体構成を示すブロック図である。

【図２３】従来例の文書検索装置の全体構成を示すブロック図である。

【符号の説明】

１　文書入力手段
２　インデクス作成手段
３　インデクス記憶手段
４　書式入力手段
５　書式記憶手段
６　検索情報作成手段
７　検索情報記憶手段
８　検索要求入力手段

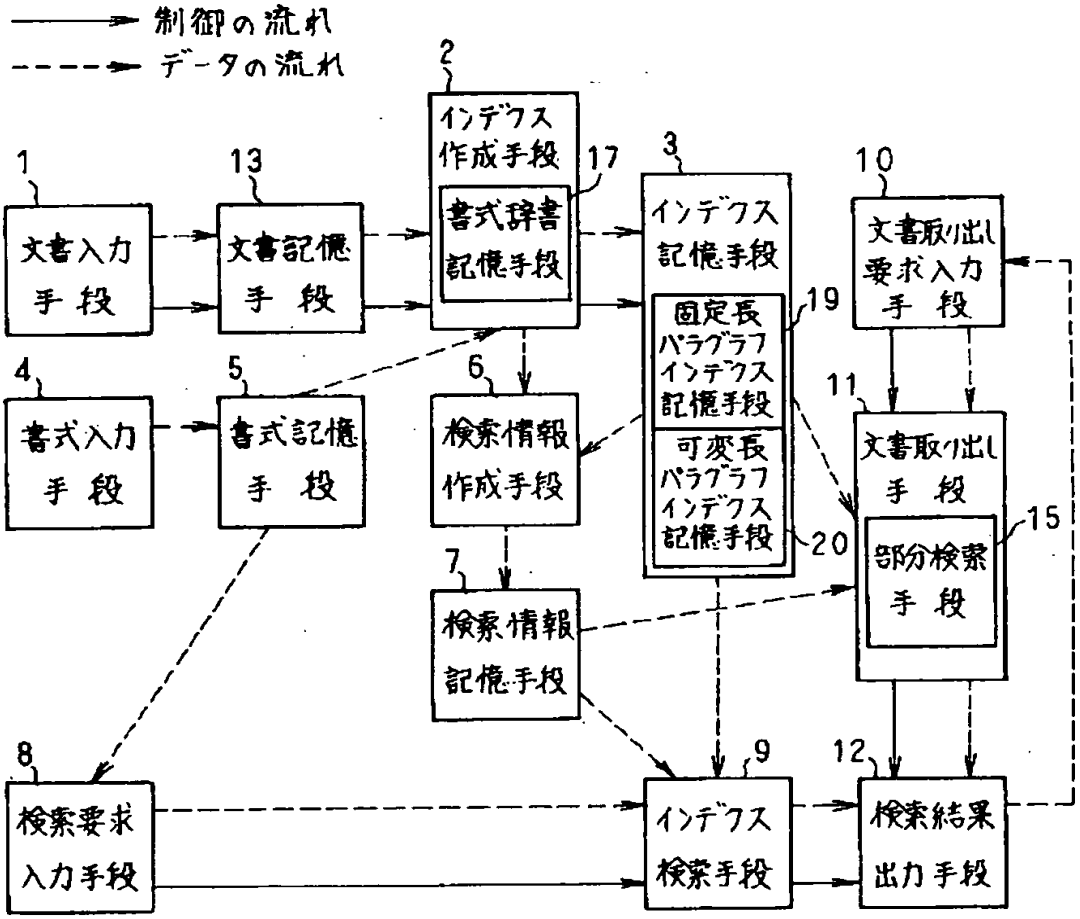
15

９　インデクス検索手段
１０　文書取り出し要求入力手段
１１　文書取り出し手段
１２　検索結果出力手段
１３　文書記憶手段

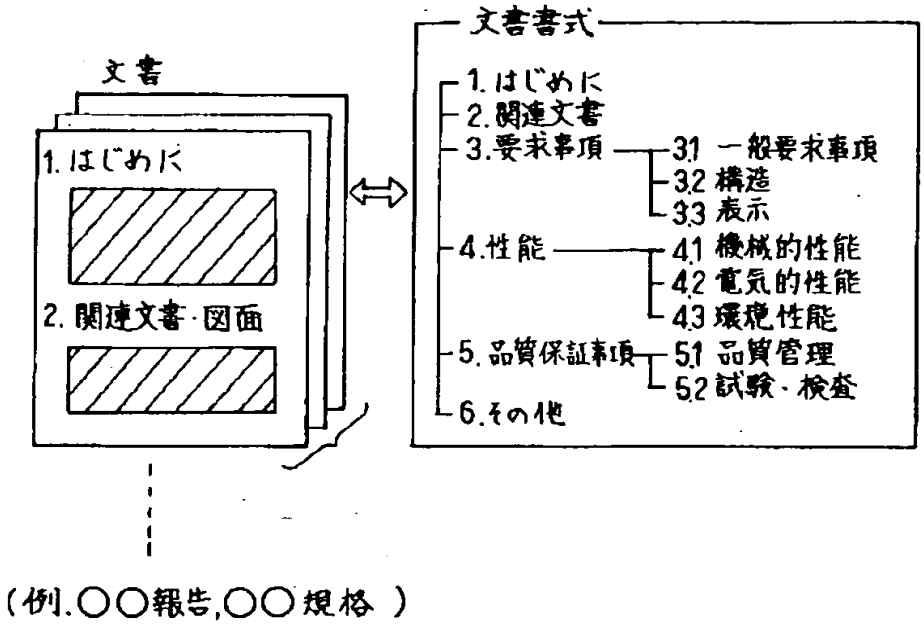
16

１５　部分検索手段
１７　書式辞書記憶手段
１９　固定長パラグラフインデクス記憶手段
２０　可変パラグラフインデクス記憶手段

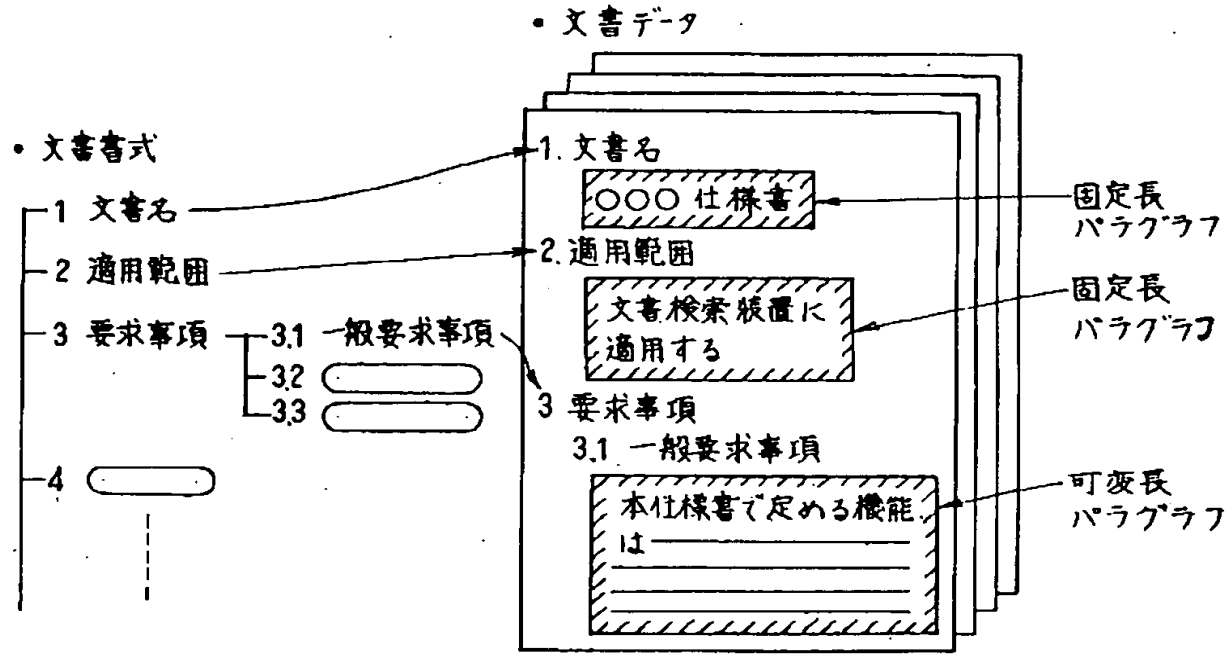
【図１】



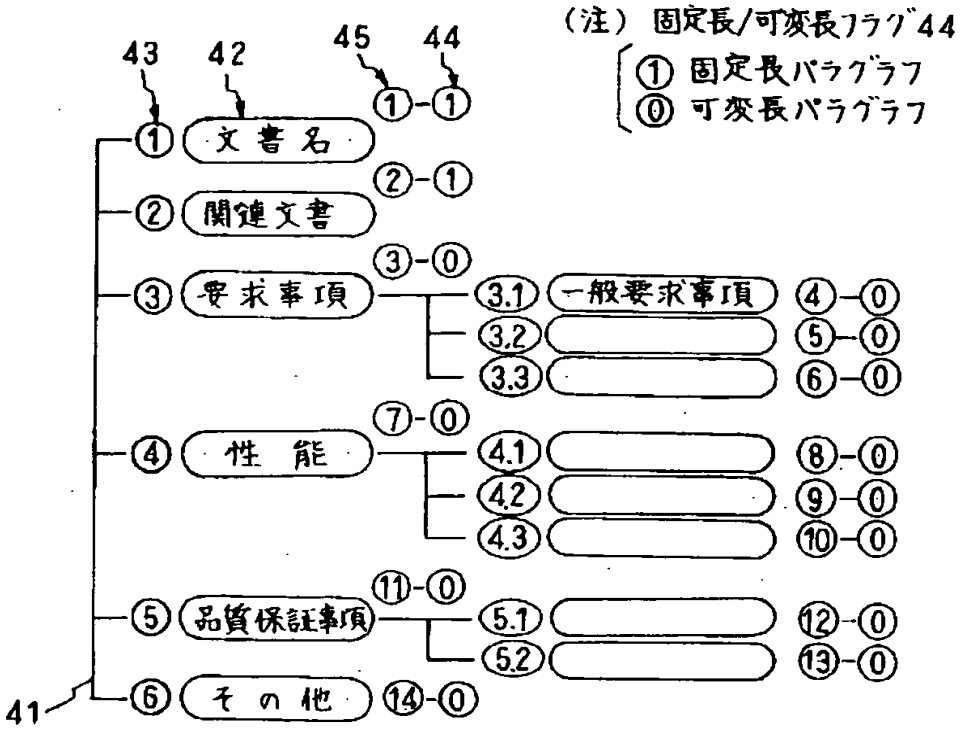
【図2】


文書
1. はじめに
2. 関連文書・図面
（例.○○報告,○○規格）
文書書式
1. はじめに
2. 関連文書
3. 要求事項
3.1 一般要求事項
3.2 構造
3.3 表示
4. 性能
4.1 機械的性能
4.2 電気的性能
4.3 環境性能
5. 品質保証事項
5.1 品質管理
5.2 試験・検査
6. その他

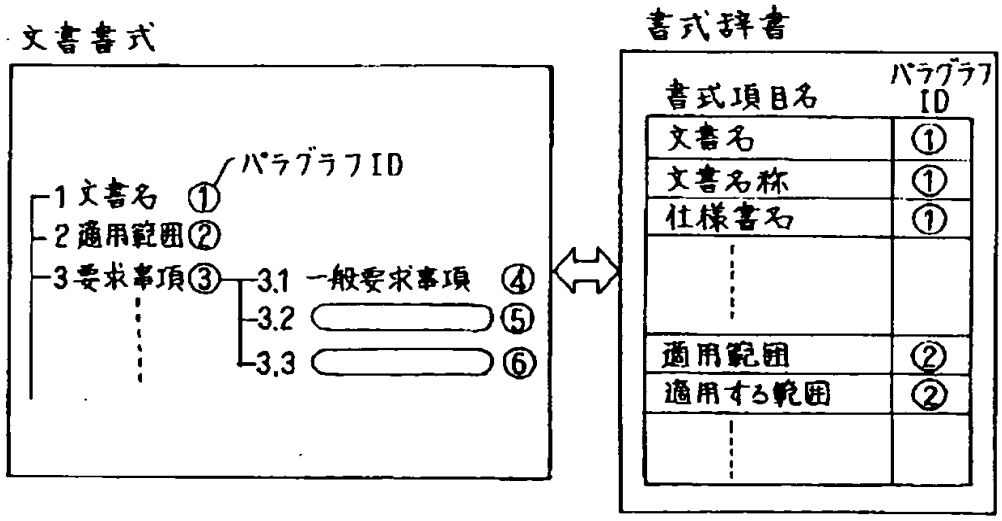
【図3】


• 文書データ
• 文書書式
1 文書名
2 適用範囲
3 要求事項
3.1 一般要求事項
3.2
3.3
4
1. 文書名
○○○ 仕様書
2. 適用範囲
文書検索装置に適用する
3 要求事項
3.1 一般要求事項
本仕様書で定める機能は
固定長パラグラフ
固定長パラグラフ
可変長パラグラフ

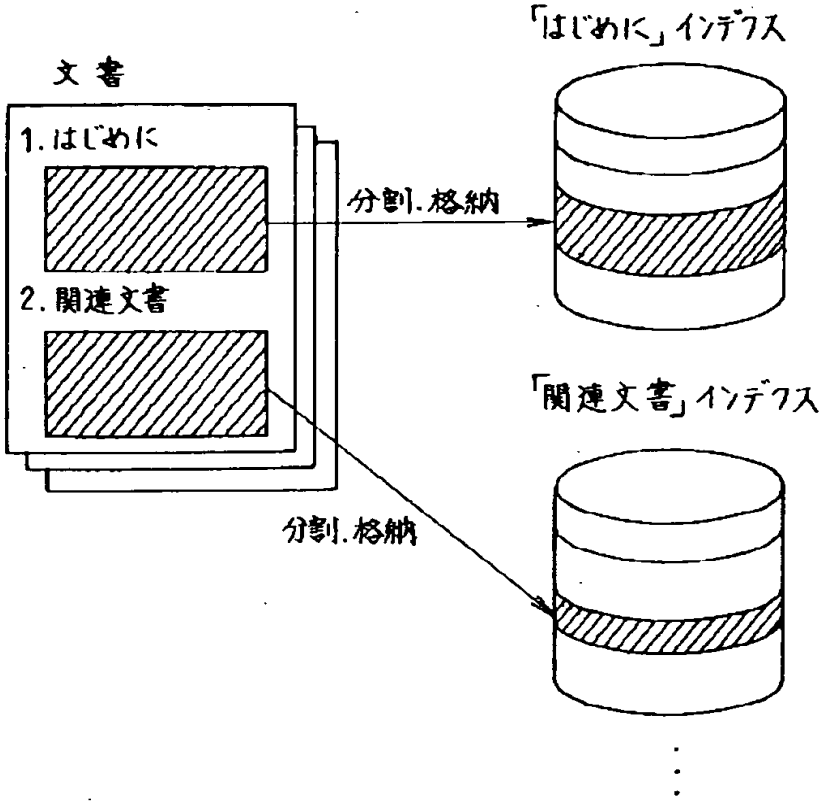
【図4】


（注） 固定長/可変長フラグ44
① 固定長パラグラフ
⓪ 可変長パラグラフ
43
42
45
44
41
① 文書名
② 関連文書
③ 要求事項
3.1 一般要求事項
3.2
3.3
④ 性能
4.1
4.2
4.3
⑤ 品質保証事項
5.1
5.2
⑥ その他
①-①
②-①
③-⓪
④-⓪
⑤-⓪
⑥-⓪
⑦-⓪
⑧-⓪
⑨-⓪
⑩-⓪
⑪-⓪
⑫-⓪
⑬-⓪
⑭-⓪

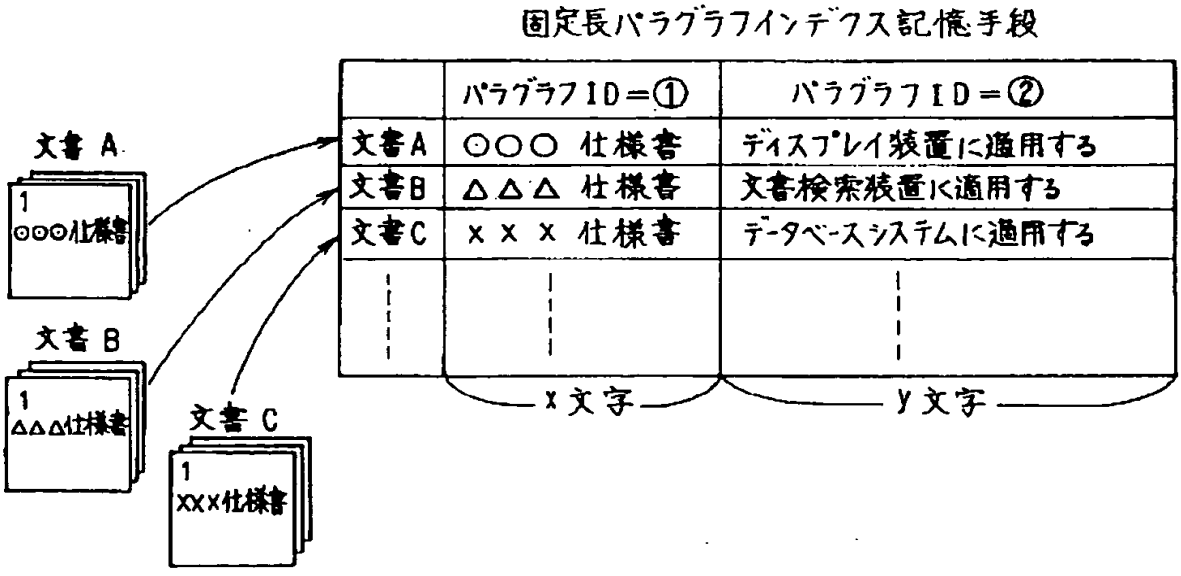
【図11】


文書書式
パラグラフID
1 文書名 ①
2 適用範囲 ②
3 要求事項 ③
3.1 一般要求事項 ④
3.2 ⑤
3.3 ⑥
書式辞書
書式項目名
パラグラフID
文書名 ①
文書名称 ①
仕様書名 ①
適用範囲 ②
適用する範囲 ②

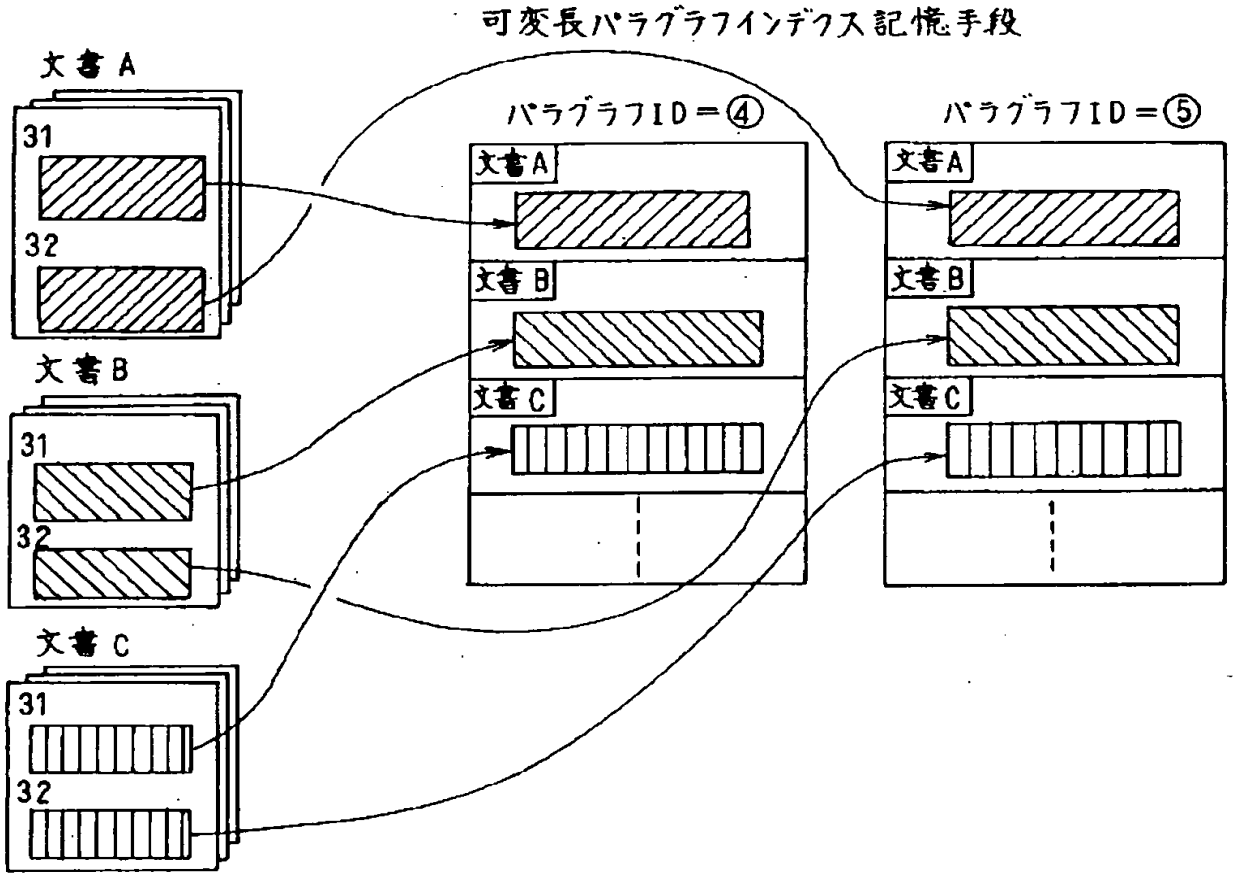
【図５】


「はじめに」インデクス
文書
1.はじめに
分割.格納
2.関連文書
「関連文書」インデクス
分割.格納

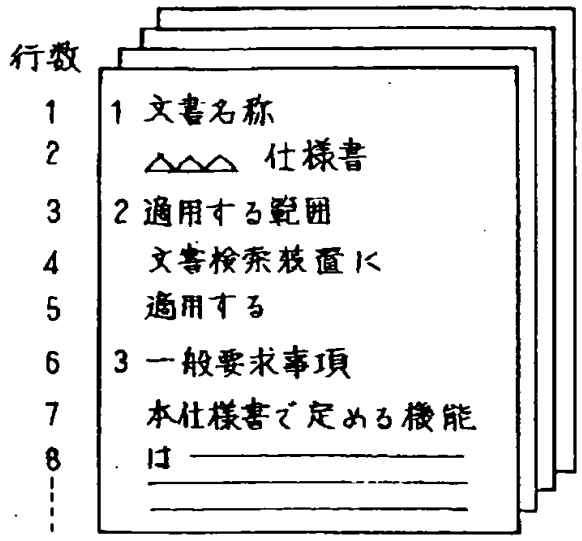
【図６】


固定長パラグラフインデクス記憶手段
パラグラフID＝①
パラグラフID＝②
文書A
○○○ 仕様書
ディスプレイ装置に適用する
文書B
△△△ 仕様書
文書検索装置に適用する
文書C
ｘｘｘ 仕様書
データベースシステムに適用する
x文字
y文字
文書 A
1
○○○仕様書
文書 B
1
△△△仕様書
文書 C
1
xxx仕様書

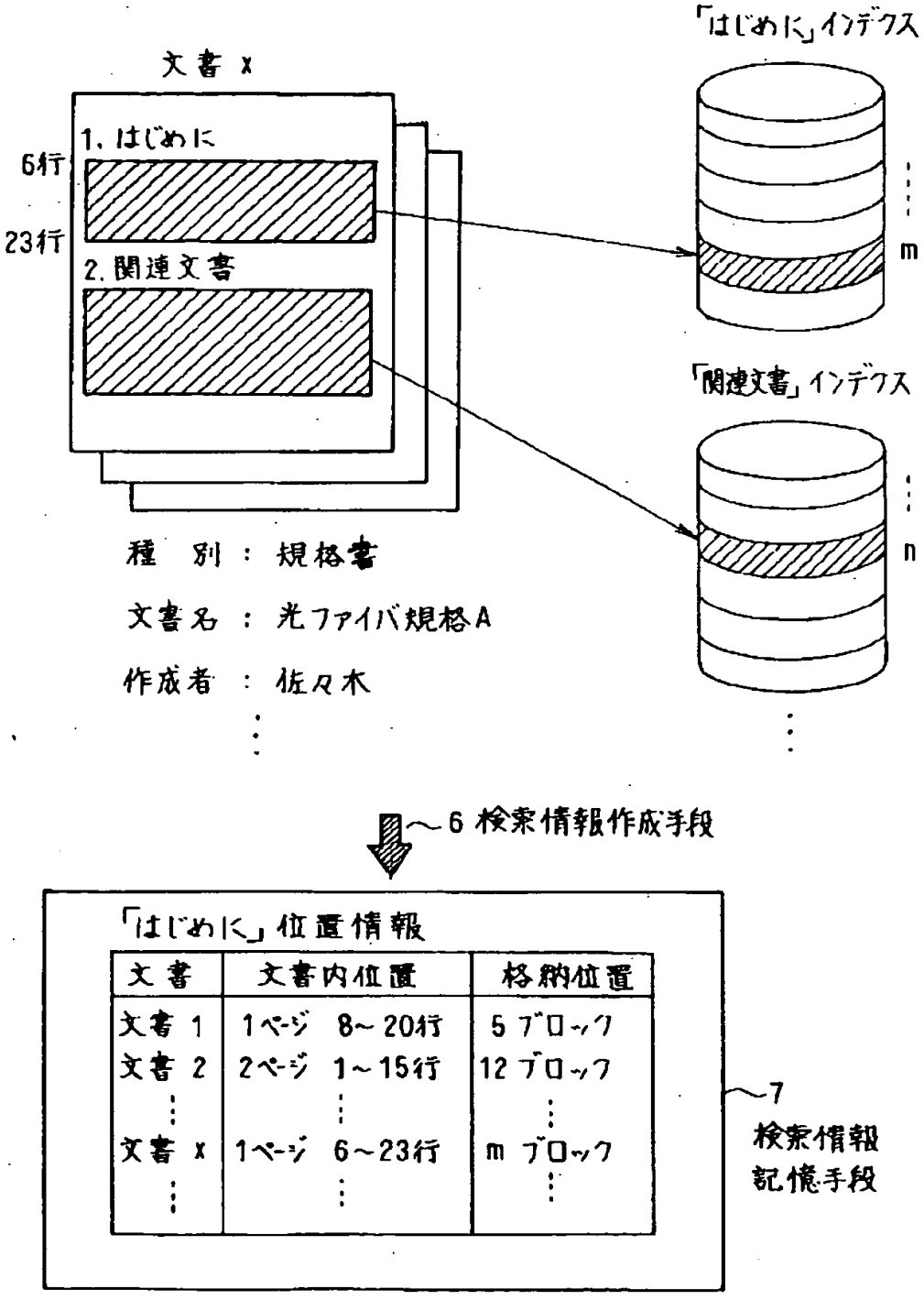
【図７】


可変長パラグラフインデクス記憶手段
文書A
31
32
文書B
31
32
文書C
31
32
パラグラフID＝④
文書A
文書B
文書C
パラグラフID＝⑤
文書A
文書B
文書C

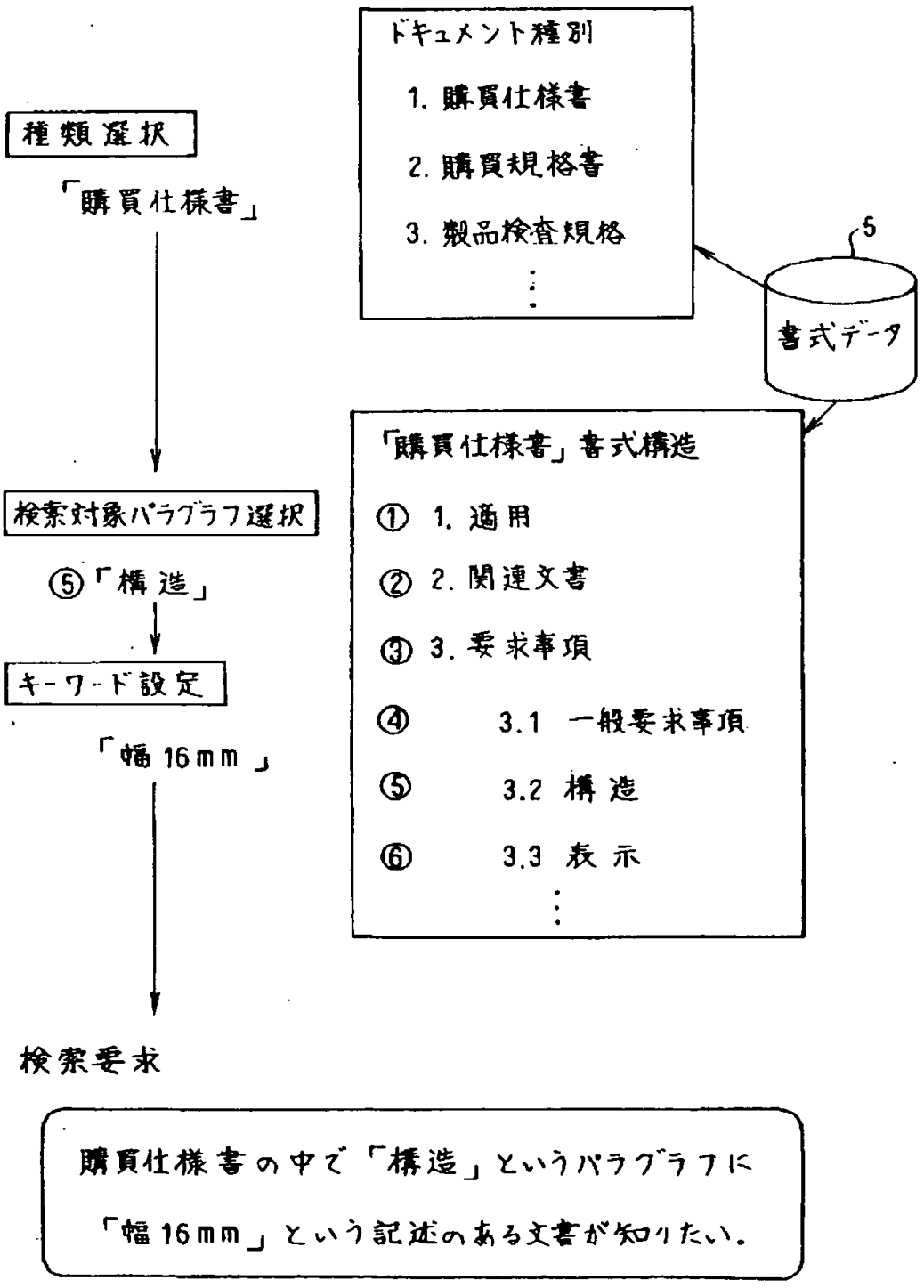
【図１４】


行数
1
2
3
4
5
6
7
8
1 文書名称
△△△ 仕様書
2 適用する範囲
文書検索装置に
適用する
3 一般要求事項
本仕様書で定める機能
は

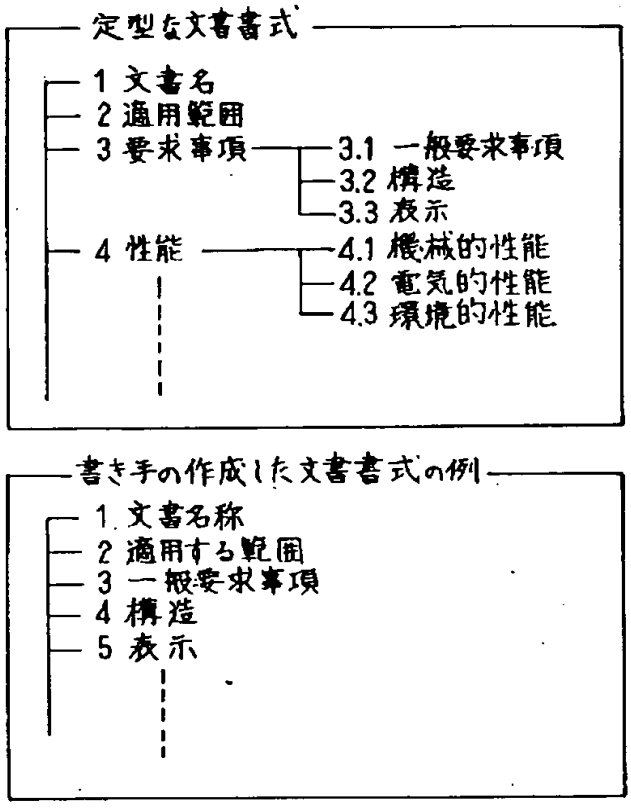
【図8】


「はじめに」インデクス
文書 x
1. はじめに
6行
23行
2. 関連文書
m
「関連文書」インデクス
n
種 別 : 規格書
文書名 : 光ファイバ規格A
作成者 : 佐々木
6 検索情報作成手段
「はじめに」位置情報
文書
文書内位置
格納位置
文書 1
1ページ 8~20行
5 ブロック
文書 2
2ページ 1~15行
12 ブロック
文書 x
1ページ 6~23行
m ブロック
7
検索情報
記憶手段

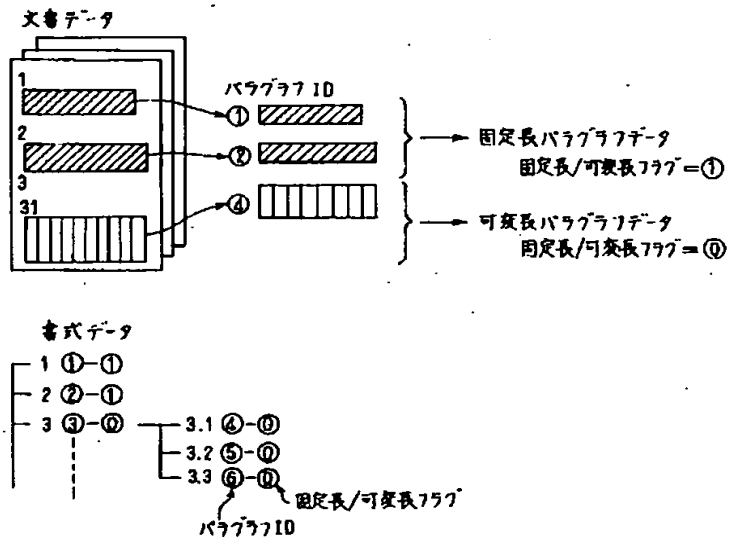
【図９】


ドキュメント種別
1. 購買仕様書
2. 購買規格書
3. 製品検査規格
5
書式データ
種類選択
「購買仕様書」
「購買仕様書」書式構造
① 1. 適用
② 2. 関連文書
③ 3. 要求事項
④ 3.1 一般要求事項
⑤ 3.2 構造
⑥ 3.3 表示
検索対象パラグラフ選択
⑤「構造」
キーワード設定
「幅 16mm」
検索要求
購買仕様書の中で「構造」というパラグラフに
「幅 16mm」という記述のある文書が知りたい。

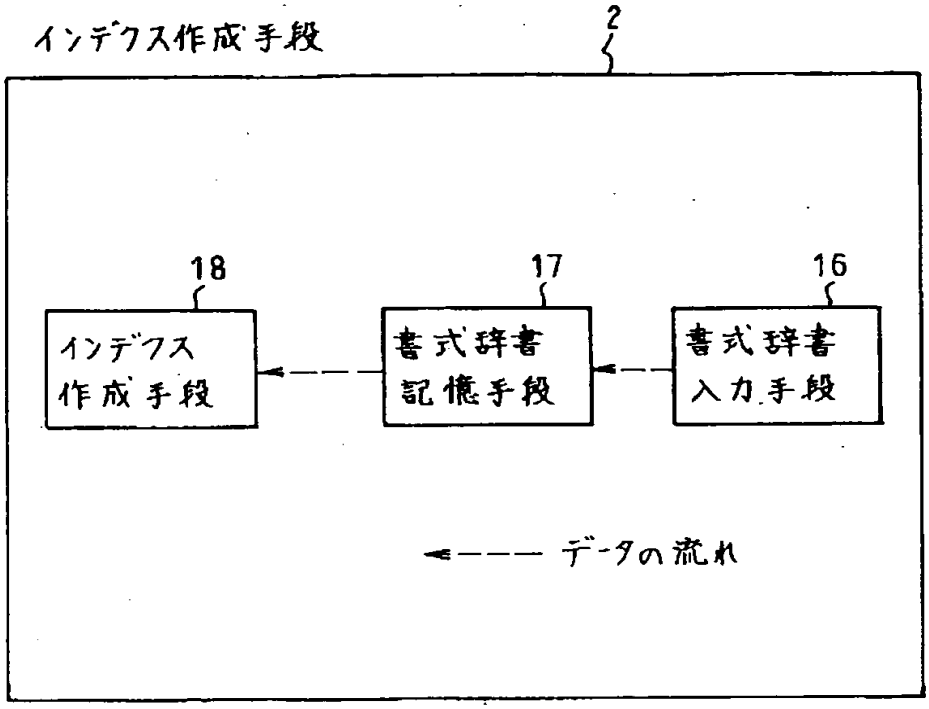
【図１０】


定型な文書書式
1 文書名
2 適用範囲
3 要求事項
3.1 一般要求事項
3.2 構造
3.3 表示
4 性能
4.1 機械的性能
4.2 電気的性能
4.3 環境的性能
書き手の作成した文書書式の例
1 文書名称
2 適用する範囲
3 一般要求事項
4 構造
5 表示

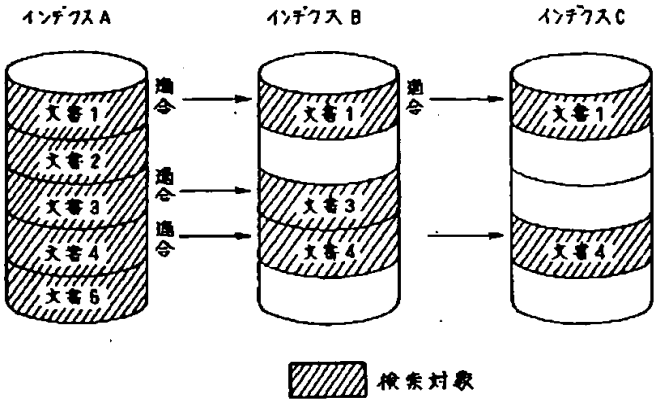
【図１５】


文書データ
1
2
3
31
パラグラフID
①
②
④
固定長パラグラフデータ
固定長/可変長フラグ=①
可変長パラグラフデータ
固定長/可変長フラグ=⓪
書式データ
1 ①-①
2 ②-①
3 ③-⓪
3.1 ④-⓪
3.2 ⑤-⓪
3.3 ⑥-⓪
固定長/可変長フラグ
パラグラフID

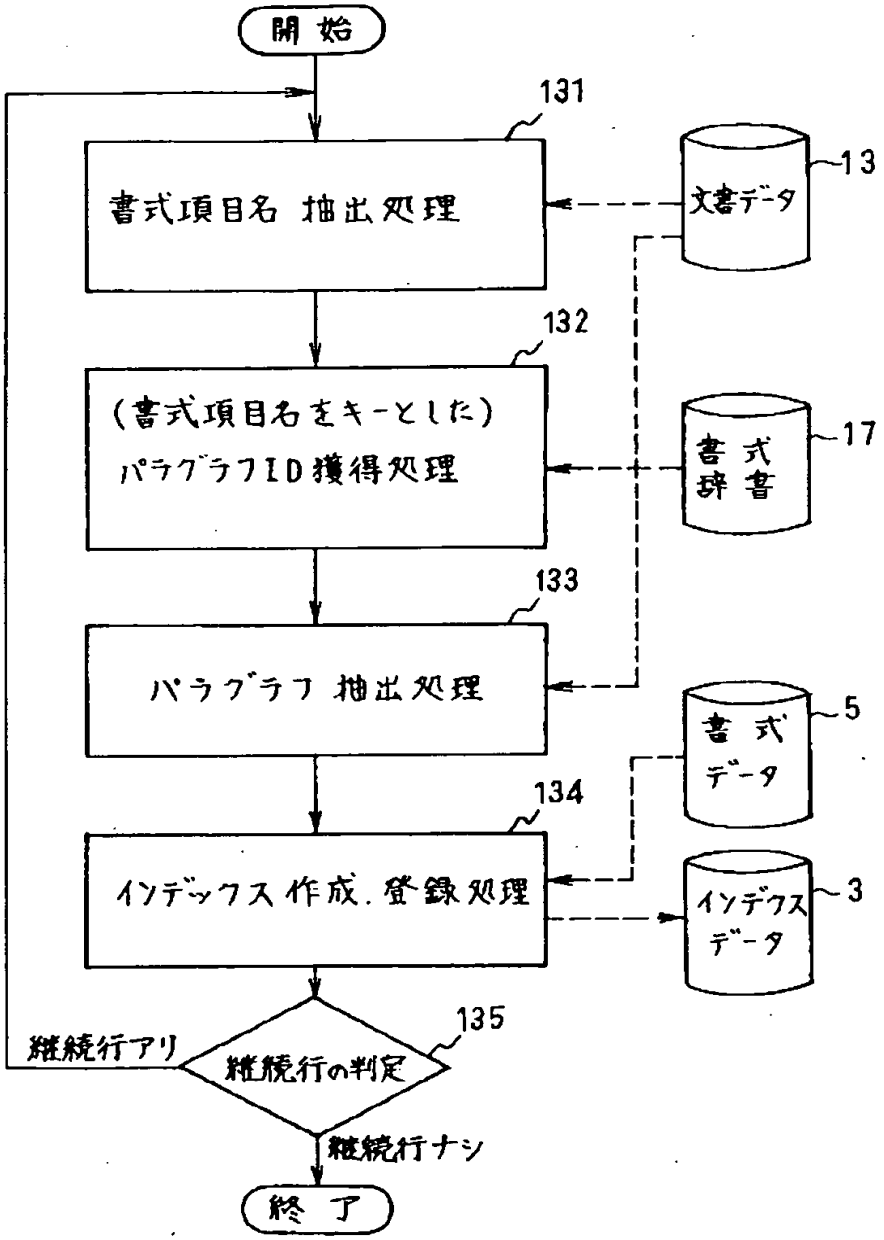
【図１２】


インデクス作成手段
2
18
17
16
インデクス
作成手段
書式辞書
記憶手段
書式辞書
入力手段
データの流れ

【図２１】


インデクスA
インデクスB
インデクスC
文書1
文書2
文書3
文書4
文書6
適合
検索対象

【図１３】


開始
131
書式項目名 抽出処理
文書データ
13
132
(書式項目名をキーとした)
パラグラフID獲得処理
書式辞書
17
133
パラグラフ 抽出処理
書式データ
5
134
インデックス作成.登録処理
インデックスデータ
3
135
継続行アリ
継続行の判定
継続行ナシ
終了

【図16】

(a) 固定長パラグラフインデクスデータ

	パラグラフID=①	パラグラフID=②
文書A	○○○仕様書	ディスプレイ装置に適用する
文書B	△△△仕様書	文書検索装置に適用する
文書C	××仕様書	データベースシステムに適用する
⋮	⋮	⋮

(b) 可変長パラグラフインデクスデータ

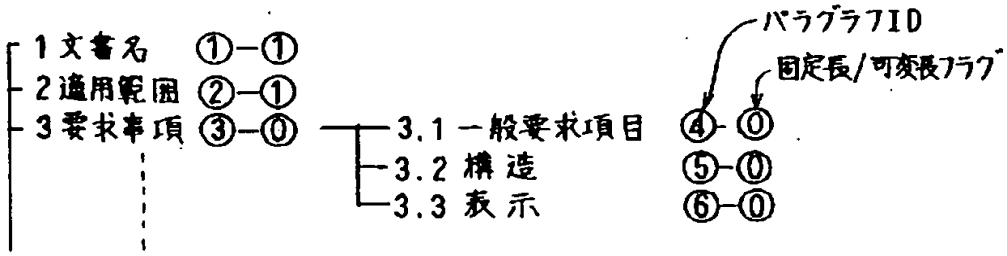
パラグラフID=④

文書A
本仕様書では、画像データを表示する－－－－
文書B
本仕様書ではマルチメディア文書を－－－－－
文書C
本仕様書ではマルチメディアデータを－－－－－
⋮

パラグラフデータID=⑤

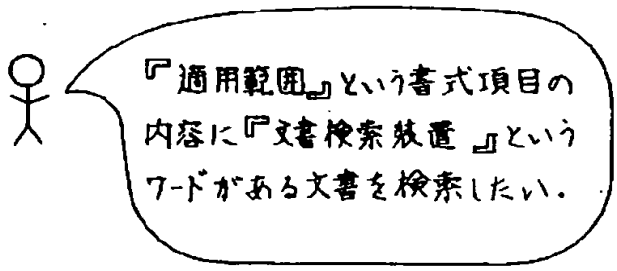
文書A
外形寸法は、たて950mm よこ300mm－－－－－
文書B
サーバマシンのデータは1GBのディスク－－－－
文書C
データとしては画像、図形、文字－－－－－－
⋮

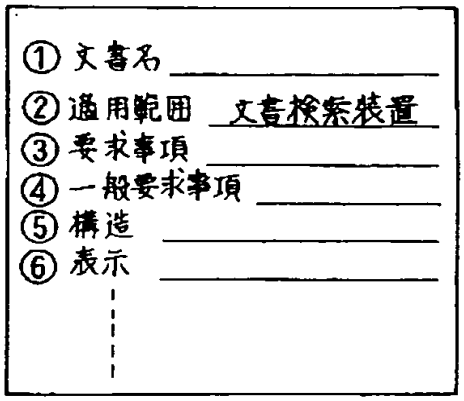
(c) 書式データ



【図17】

（a）検索要求





（下線部にキーワードの入力可）

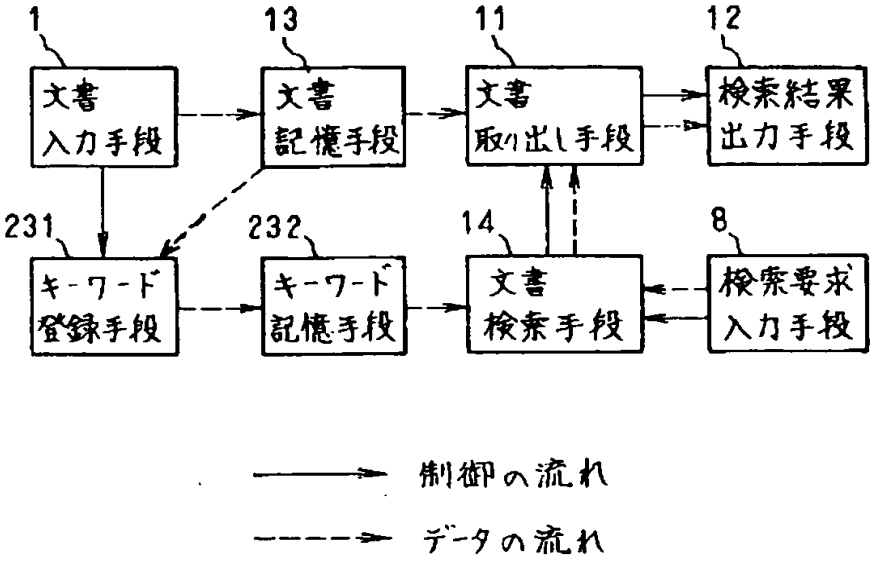
（b）検索条件

パラグラフID		キーワード		固定長／可変長フラグ
②	=	『文書検索装置』		①

（c）検索結果

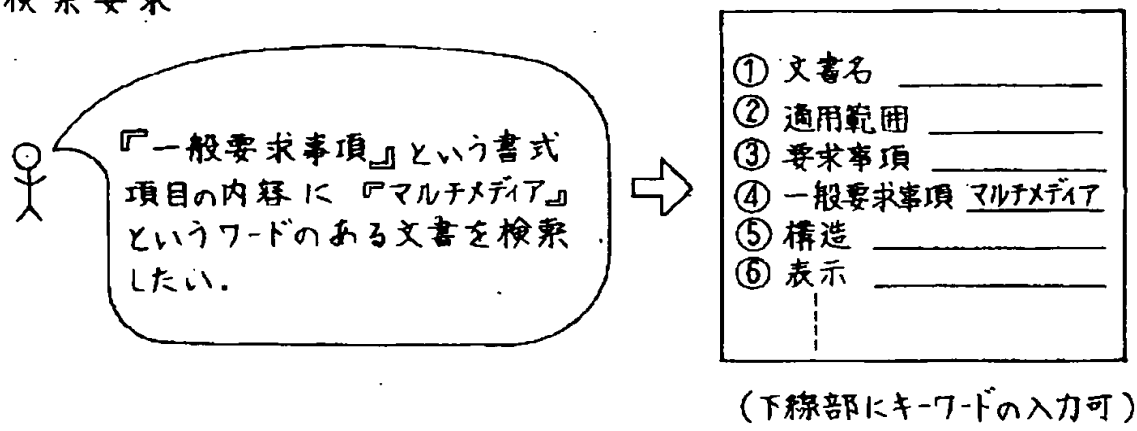
文書B

【図23】



【図１８】

(a)検索要求



（下線部にキーワードの入力可）

(b)検索条件

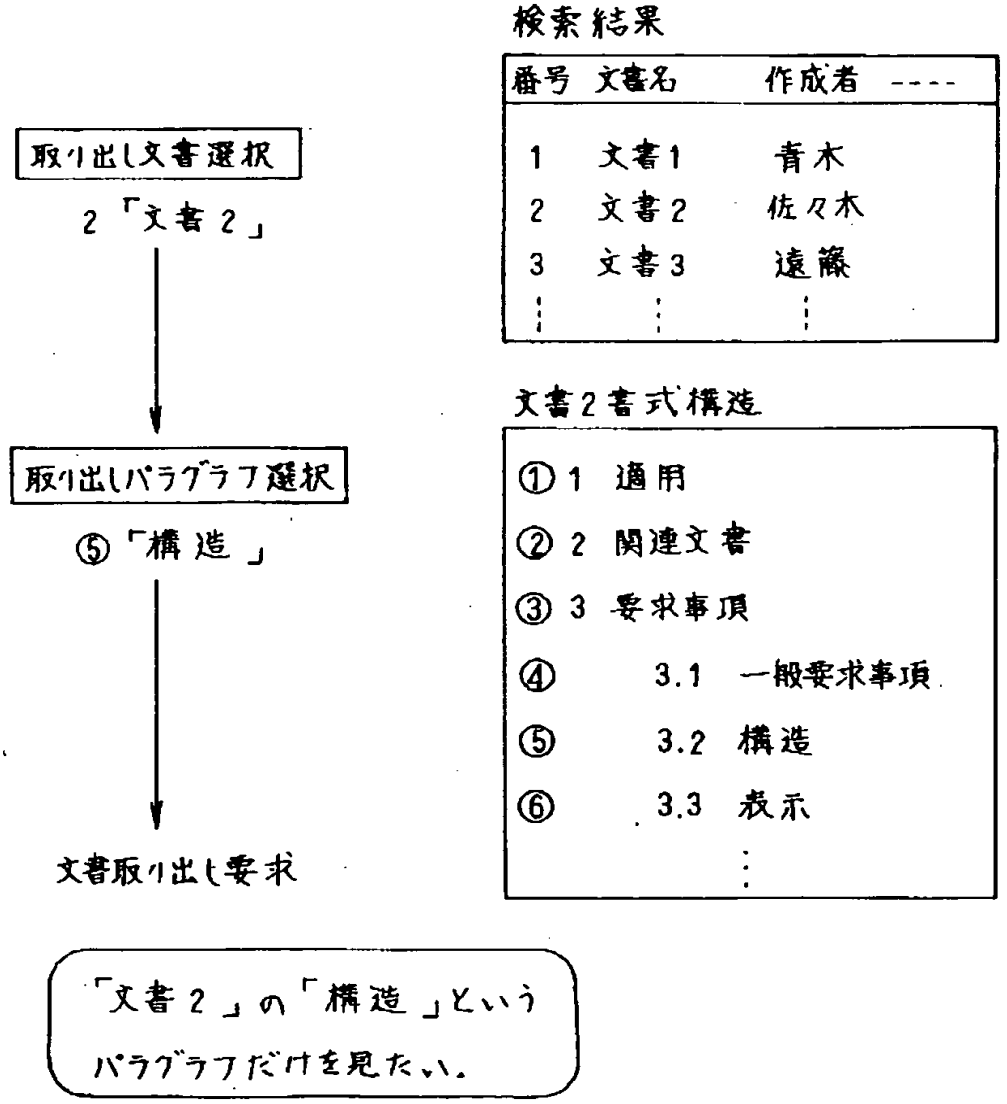
パラグラフID		キーワード	固定長/可変長フラグ
④	＝	『マルチメディア』	⓪

(c)検索結果

文書 A

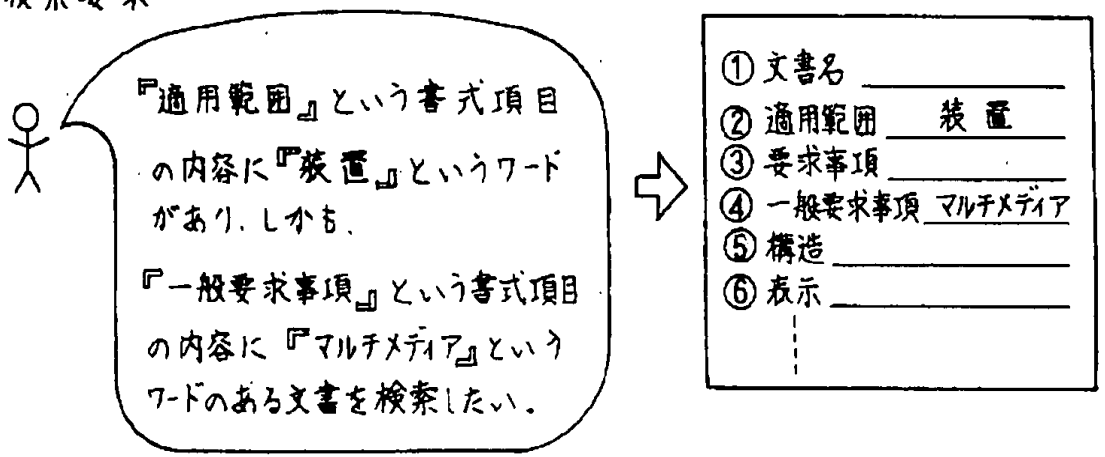
文書 B

【図１９】


検索結果
番号 文書名 作成者 ----
1 文書1 青木
2 文書2 佐々木
3 文書3 遠藤
取り出し文書選択
2「文書２」
取り出しパラグラフ選択
⑤「構造」
文書取り出し要求
文書2書式構造
① 1 適用
② 2 関連文書
③ 3 要求事項
④ 3.1 一般要求事項
⑤ 3.2 構造
⑥ 3.3 表示
「文書２」の「構造」という
パラグラフだけを見たい。

【図２０】

(a)検索要求



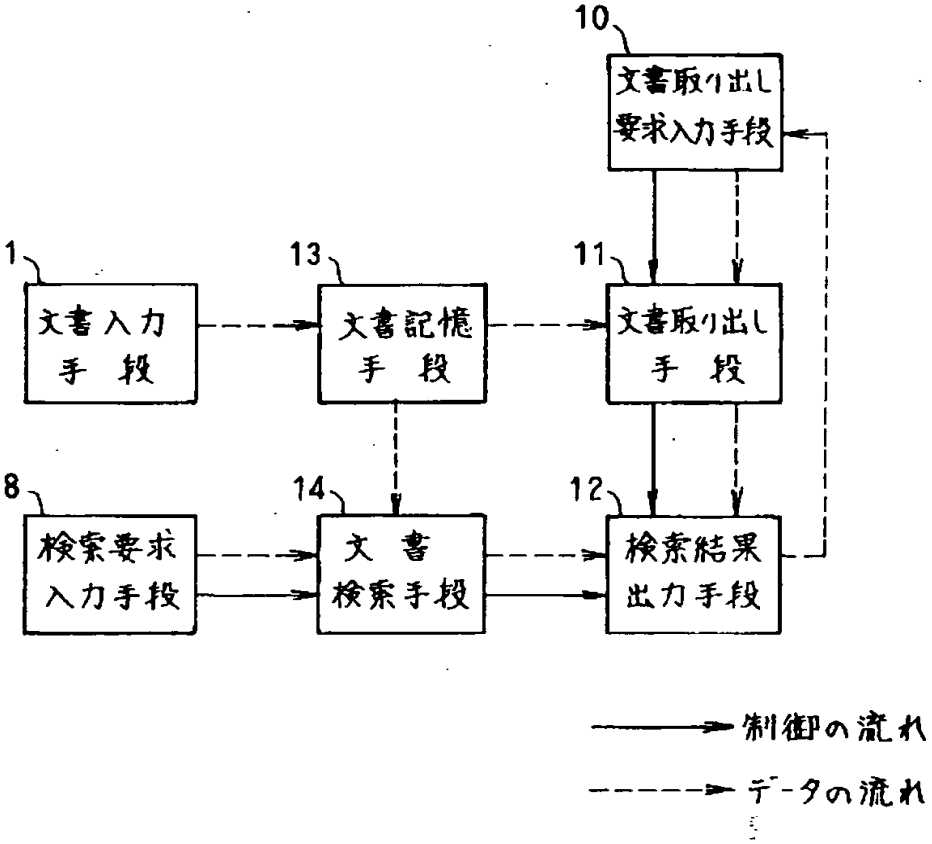
(b)検索条件

パラグラフID		キーワード	固定長/可変長フラグ
②	=	『装 置』	①
and			
④	=	『マルチメデア』	⓪

(c)検索結果

文書 B

【図２２】


10
文書取り出し
要求入力手段
1
文書入力
手段
13
文書記憶
手段
11
文書取り出し
手段
8
検索要求
入力手段
14
文書
検索手段
12
検索結果
出力手段
制御の流れ
データの流れ